**R Project 1**

**Covid-19 Data Analysis**

rm(list=ls()) # removes all variables stored previously

library(Hmisc) # import

> rm(list=ls()) # removes all variables stored previously

> library(Hmisc) # import

data <- read.csv("C:/Users/LENOVO/Downloads/COVID19\_line\_list\_data(1).csv")

describe(data) # Hmisc command

> data <- read.csv("C:/Users/LENOVO/Downloads/COVID19\_line\_list\_data(1).csv")

> describe(data) # Hmisc command data

# cleaned up death column

data$death\_dummy <- as.integer(data$death != 0)

# death rate

sum(data$death\_dummy) / nrow(data)

> # cleaned up death column

> data$death\_dummy <- as.integer(data$death != 0)

> # death rate

> sum(data$death\_dummy) / nrow(data)

[1] 0.05806452

# AGE

# claim: people who die are older

dead = subset(data, death\_dummy == 1)

alive = subset(data, death\_dummy == 0)

mean(dead$age, na.rm = TRUE)

mean(alive$age, na.rm = TRUE)

> # AGE

> # claim: people who die are older

> dead = subset(data, death\_dummy == 1)

> alive = subset(data, death\_dummy == 0)

> mean(dead$age, na.rm = TRUE)

[1] 68.58621

> mean(alive$age, na.rm = TRUE)

[1] 48.07229

# is this statistically significant?

t.test(alive$age, dead$age, alternative="two.sided", conf.level = 0.99)

# normally, if p-value < 0.05, we reject null hypothesis

# here, p-value ~ 0, so we reject the null hypothesis and

# conclude that this is statistically significant

data: alive$age and dead$age

t = -10.839, df = 72.234, p-value <

2.2e-16

alternative hypothesis: true difference in means is not equal to 0

99 percent confidence interval:

-25.52122 -15.50661

sample estimates:

mean of x mean of y

48.07229 68.58621

> # normally, if p-value < 0.05, we reject null hypothesis

> # here, p-value ~ 0, so we reject the null hypothesis and

> # conclude that this is statistically significant

# GENDER

# claim: gender has no effect

men = subset(data, gender == "male")

women = subset(data, gender == "female")

mean(men$death\_dummy, na.rm = TRUE) #8.5%!

mean(women$death\_dummy, na.rm = TRUE) #3.7%

> # GENDER

> # claim: gender has no effect

> men = subset(data, gender == "male")

> women = subset(data, gender == "female")

> mean(men$death\_dummy, na.rm = TRUE) #8.5%!

[1] 0.08461538

> mean(women$death\_dummy, na.rm = TRUE) #3.7%

[1] 0.03664921

# is this statistically significant?

t.test(men$death\_dummy, women$death\_dummy, alternative="two.sided", conf.level = 0.99)

# 99% confidence: men have from 0.8% to 8.8% higher chance

# of dying.

# p-value = 0.002 < 0.05, so this is statistically

# significant

alternative hypothesis: true difference in means is not equal to 0

99 percent confidence interval:

0.007817675 0.088114665

sample estimates:

mean of x mean of y

0.08461538 0.03664921

> # 99% confidence: men have from 0.8% to 8.8% higher chance

> # of dying.

> # p-value = 0.002 < 0.05, so this is statistically

> # significant